High resolution mapping of cross-over events in cattle using NGS data

N. Kadri[1], C.Harland[1], C. Charlier[1], L.Karim[2], N. Cambisano[2], M. Deckers[2], E. Mullaart[3], W. Coppieters[2], M.Georges[1], T.Druet[1]

[1] Unit of Animal genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liege,Belgium
nk.kadri@ulg.ac.be (Corresponding Author)
[2] GIGA Genomics Platform, University of Liège, Belgium
[3] CRV, Arnhem, The Netherlands

Keywords: recombination, cattle, NGS

Summary

Homologous recombination plays an important role in proper segregation of homologues in the first meiotic division. Failure in proper segregation results in aneuploidy, which is a leading cause for pregnancy loss in humans. Recently, global recombination rate has been studied in large cattle populations genotyped with SNP arrays (~50K). However, the fine-scale resolution of these studies remained limited as a result of the relatively low marker density. Here we report high-resolution mapping of cross-over (CO) events in a cattle pedigree using whole genome sequence data. We carry out an extensive cleaning of our sequence data to remove errors (errors in the genome build, sequencing errors and presence of CNVs) that dramatically inflate CO counts. Using ~5 million high quality sequence variants we identify 3,880 CO events in 155 male gametes and 3,088 CO events in 124 female gametes. The median resolution of the identified COs was 34 kb with about 75% of the events mapped to an interval less than 100 kb. The male and female map lengths were estimated at 27.5 M and 23.8 M respectively. Consistent with previous studies in cattle, we find higher recombination rate in males and higher frequency of COs at chromosome ends. Interestingly, compared to the map lengths estimated from SNP chip we find an increase of 3.7 and 2.7 M in male and female maps respectively. Despite the cleaning efforts, we cannot determine at this time whether the increased in map lengths correspond to CO missed with genotyping arrays, to spurious CO identified with NGS data (due to unidentified sources of errors) or both.

Keywords: recombination, cattle, NGS

Introduction

Homologous recombination plays an important role in gametogenesis. It provides tension necessary for the meiotic spindle apparatus for the proper segregation of homologues in meiosis I. The right number and placement of cross-overs on the tetrad are thus necessary for the synthesis of viable gametes. Failure in proper segregation results in aneuploidy, which is a leading cause of pregnancy loss in humans (Hassold & Hunt 2001). In addition, recombination plays a role in evolution as it creates new allele combinations and increases the genetic diversity for natural selection to act upon.

Recombination rate has been recently studied in cattle using genotyping array data (e.g.,

50K genotyping array). These studies reported the genome-wide recombination rates (GRR), their heritability and identified genetic determinants accounting for variation in GRR (Sandor et al., 2012, Ma et al., 2015, Kadri et al., 2016). However, a higher density of markers is necessary to study the fine-scale patterns of recombination, especially to identify recombination hotspots and to search for motifs associated with cross-overs (CO). Here we use whole-genome sequence (WGS) data from a large dairy cattle pedigree including 743 sequenced animals to identify COs at high resolution.

Material and methods

743 animals belonging to an extended pedigree were sequenced in the DAMONA project (Harland et al., 2017). For this study we selected 266 animals sequenced at depth 15x or higher. The animals belonged to either a 3 generational pedigree (grand-parent / parent/ offspring) or to a half-sib family with at least 3 offspring and a parent genotyped (see Table1). Parent and offspring were phased following Mendelian rules using the grand-parent and parent data respectively. Linkage information (3 half-sibs or more) was used to phase the parent when the grand-parent was not available and to improve the phasing of parents in the half-sibs families. Stringent filtering rules (SNP behaving like true Mendelian variants; see (Kadri et al., 2016) for more details) were applied to select 5,366,864 bi-allelic variants segregating in multiple WGS datasets (1000 bulls genome, Belgian Blue cattle, Holstein-Jersey population) for further analyses. Haplotype reconstruction and CO identification was performed using LINKPHASE3 (Druet & Georges 2015) .

Results and Discussions

Identifying sources of inflated recombination rates

Different types of errors can cause identification of spurious CO and inflate estimated GRR. First, errors in the UMD3.1 genome assembly (segment mapped to an incorrect position) can potentially generate double COs. Without map corrections; the sex averaged genetic map length was 85M. In order to identify the misplaced segments in the genome build, we utilized a “clean” 50k SNP map (Druet & Georges 2015, Kadri al., 2016) and calculated squared-correlations (r^2) between inheritance vectors (determining the inherited homologs) estimated at every sequence position and the 2 flanking SNP positions (from the 50K). The boundaries of these putative map errors (with reduced r^2) were identified by combining the r^2 values with map confidence scores obtained with LINKPHASE3 (Figure 1). In total, we identified 164 possibly misplaced segments (representing a total of 16 Mb) and removed them from our data set. After removal of these putative map errors, the average genome length dropped from 91M to 33M.

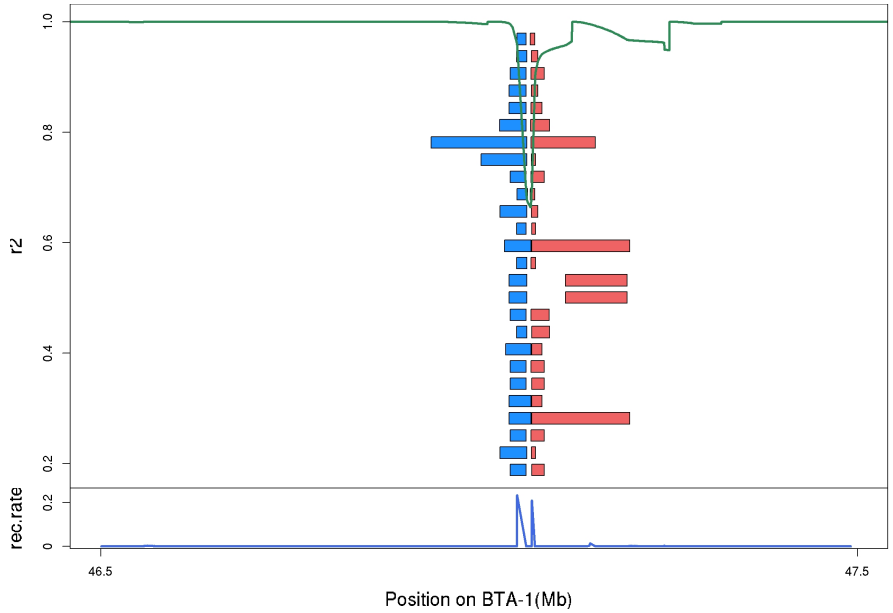


Figure 1. Illustration showing identification of map/build error.
The green line shows the r^2 between a variant from a reference map (based on the SNP chip) and the sequence variants (see methods). The blue and red bars show double cross-overs in 26 gametes resulting from the misplaced segment. Blue line shows the recombination rate between consecutive sequence variants estimated using EM algorithm implemented in LINKPHASE3.

In addition to putative map errors, causing errors in all families, we also observed spurious CO (e.g., excess CO in a small region) specific to certain families. These might typically be the consequence of genotyping errors. Although, LINKPHASE3 is robust to genotyping errors, the presence of clustered genotyping errors can create problems. We previously determined that LINKPHASE3 was more efficient when genotype inconsistencies were previously cleaned (e.g., parent-offspring pairs having opposite homozygote genotypes for a marker). Clustered genotyping inconstancies might result from deletions (individuals incorrectly called homozygotes). On the raw WGS data, removal of variants presenting genotype incompatibilities drastically reduced the number of CO per meiosis from 962 to 102.

Similarly, presence of duplications generates genotyping errors but these do not generally cause genotype incompatibilities (since they generally cause excess heterozygosity). We observed the presence of such Copy Number Variants (CNVs) associated with spurious CO (Figure 2). We developed a tool to identify CNVs based on allelic depth and allele frequencies and removed those prior to phasing.

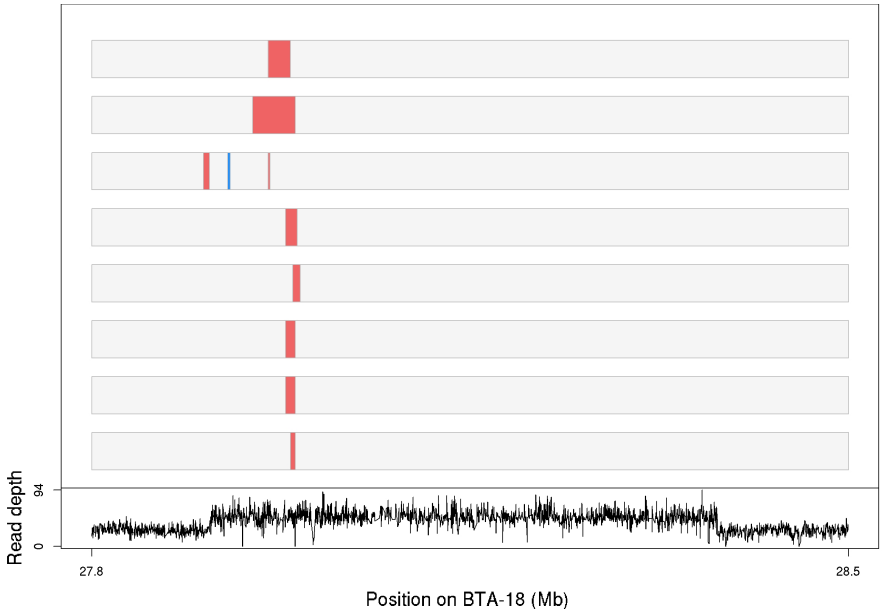
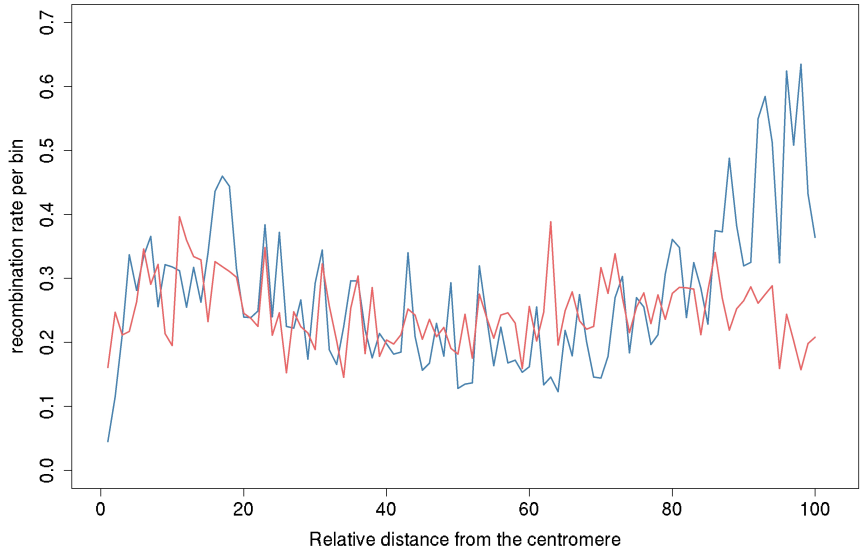


Figure 2. Duplication and spurious CO on BTA-18
The red and blue bars show CO intervals in 7 gametes from a single parent. The lower panel shows the increased read depth of the parent in the CNV region.

Recombination rate and distribution of COs

After removing the identified sources of errors, we realized a high-resolution mapping of CO in cattle using NGS data. A total of 7535 COs were identified in 279 gametes (see Table 1) with an estimated map lengths of 27.5 M and 24.9 M in males and females respectively. Consistent with previous studies in cattle, we find higher recombination rate in males with more COs at the chromosome ends (Figure 3) (Ma et al., 2015, Kadri et al., 2016). Interestingly, compared to estimates from the 50K SNP chip, we observe an increase of 3.7M (27.47 from 23.75) and 2.72M (24.86 from 22.14) in male and female map lengths respectively. As explained earlier, we have carried out extensive cleaning of the genotype data removing (i) putative errors in the genome assembly (ii) Mendelian inconsistencies and (iii) incorrect genotypes associated with CNVs. All together, these cleaning measures resulted in a ~3.5 fold reduction in sex averaged map length from 90.5 M to 25.6 M. Despite the cleaning efforts, we cannot determine at this time whether the increased map lengths correspond to CO missed with genotyping arrays, or to spurious CO identified with NGS data (due to unidentified sources of errors) or both.



-Figure 3. Distribution of CO events along the chromosomes in females (red) and males (blue) pooled across 29 autosomes.

With ~5 million high quality variants, NGS data provided a ~160 fold increase in marker density compared to 50K SNP chip. This allowed us to map CO events to shorter intervals (Figure 4). The median resolution of identified COs improved from 1,075 Kb to 34 Kb, with about 75% of COs mapped within an interval of 100 Kb (Figure 4). The resolution can be further improved to a median resolution of 13.2 Kb if all available (~13 million) variants are used to refine the COs identified with high quality variants.

Table 1. Number of parents, gametes and cross-overs

	With grand parent			Without grand parent			
	Parents	Gametes	Crossovers	Parents	Gametes	Crossovers	GRR[1]
Male	21	51	1,401	16	104	2,479	27.47
Female	61	112	2,784	2	12	304	23.75
All	82	163	4,185	18	116	2,783	

[1] GRR estimates are made from gametes with sequenced grandparent

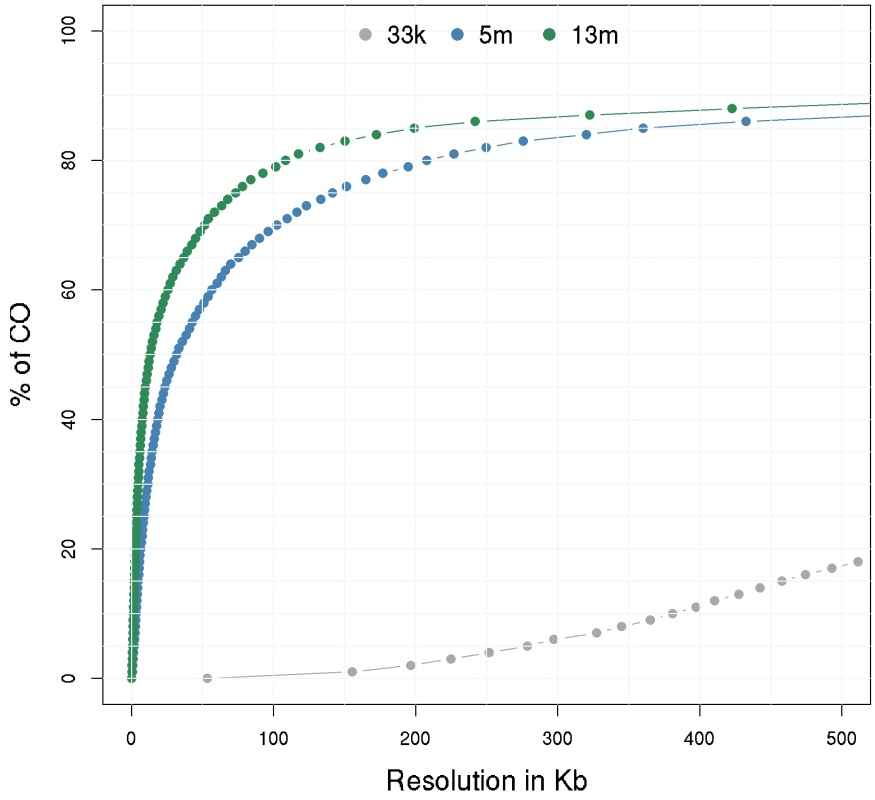


Figure 4. Resolution of identified cross-overs.
The resolution of cross-overs identified using ~ 33K markers from the SNP chip (gray), ~5 million sequence variants (blue) and ~13 million variants (green).

List of References

Druet, T. and M. Georges (2015). "LINKPHASE3: an improved pedigree-based phasing algorithm robust to genotyping and map errors." Bioinformatics **31**(10): 1677-1679.
Harland, C., C. Charlier, L. Karim, N. Cambisano, M. Deckers, M. Mni, E. Mullaart, W. Coppieters and M. Georges (2017). "Frequency of mosaicism points towards mutation-prone early cleavage cell divisions." bioRxiv: 079863.
Hassold, T. and P. Hunt (2001). "To err (meiotically) is human: the genesis of human aneuploidy." Nature reviews. Genetics **2**(4): 280.
Kadri, N. K., C. Harland, P. Faux, N. Cambisano, L. Karim, W. Coppieters, S. Fritz, E. Mullaart, D. Baurain and D. Boichard (2016). "Coding and noncoding variants in HFM1, MLH3, MSH4, MSH5, RNF212, and RNF212B affect recombination rate in cattle." Genome research **26**(10): 1323-1332.
Ma, L., J. R. O'Connell, P. M. VanRaden, B. Shen, A. Padhi, C. Sun, D. M. Bickhart, J. B. Cole, D. J. Null and G. E. Liu (2015). "Cattle sex-specific recombination and genetic control from a large pedigree analysis." PLoS genetics **11**(11): e1005387.
Sandor, C., W. Li, W. Coppieters, T. Druet, C. Charlier and M. Georges (2012). "Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle." PLoS

genetics **8**(7): e1002854.